Springer Texts in Statistics

Peter K. Dunn · Gordon K. Smyth

Generalized Linear Models With Examples in R

Springer

Springer Texts in Statistics

More information about this series at http://www.springer.com/series/417

Peter K. Dunn • Gordon K. Smyth

Generalized Linear Models With Examples in R

Springer

Peter K. Dunn
Faculty of Science, Health, Education
and Engineering
School of Health of Sport Science
University of the Sunshine Coast
QLD, Australia

Gordon K. Smyth
Bioinformatics Division
Walter and Eliza Hall Institute
of Medical Research
Parkville, VIC, Australia

ISSN 1431-875X ISSN 2197-4136 (electronic)
Springer Texts in Statistics
ISBN 978-1-4419-0117-0 ISBN 978-1-4419-0118-7 (eBook)
https://doi.org/10.1007/978-1-4419-0118-7

Library of Congress Control Number: 2018954737

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

To my wife Alison; our children Jessica, Emily, Jemima, Samuel, Josiah and Elijah; and my parents: Thank you for your love and support and for giving so much so I could get this far. PKD

To those who taught me about glms 40 years ago and to all the students who, in the years since, have patiently listened to me on the subject, given feedback and generally made it rewarding to be a teacher. GKS

Preface

A sophisticated analysis is wasted if the results cannot be communicated effectively to the client.
Reese [4, p. 201]

Our purpose in writing this book is to combine a good applied introduction to generalized linear models (GLMs) with a thorough explanation of the theory that is understandable from an elementary point of view.

We assume students to have basic knowledge of statistics and calculus. A working familiarity with probability, probability distributions and hypothesis testing is assumed, but a self-contained introduction to all other topics is given in the book including linear regression. The early chapters of the book give an introduction to linear regression and analysis of variance suitable for a second course in statistics. Students with more advanced backgrounds, including matrix algebra, will benefit from optional sections that give a detailed introduction to the theory and algorithms. The book can therefore be read at multiple levels. It can be read by students with only a first course in statistics, but at the same time, it contains advanced material suitable for graduate students and professionals.

The book should be appropriate for graduate students in statistics at either the masters or PhD levels. It should be also be appropriate for advanced undergraduate students taking majors in statistics in Britain or Australia. Students in psychology, biometrics and related disciplines will also benefit. In general, it is appropriate for anyone wanting a practical working knowledge of GLMs with a sound theoretical background.

R is a powerful and freely available environment for statistical computing and graphics that has become widely adopted around the world. This book includes a self-contained introduction to R (Appendix A), and use of R is integrated into the text throughout the book. This includes comprehensive R code examples and complete code for most data analyses and case studies. Detailed use of relevant R functions is described in each chapter.

A practical working knowledge of good applied statistical practice is developed through the use of real data sets and numerous case studies. This book makes almost exclusive use of real data. These data sets are collected in the R package **GLMsData** [1] (see Appendix A for instructions for obtaining this R package), which has been prepared especially for use with this book and which contains 97 data sets. Each example in the text is cross-referenced with the relevant data set so that readers can load the relevant data to follow the analysis in their own R session. Complete reproducible R code is provided with the text for most examples.

The development of the theoretical background sometimes requires more advanced mathematical techniques, including the use of matrix algebra. However, knowledge of these techniques is not required to read this book. We have ensured that readers without this knowledge can still follow the theoretical * development, by flagging the corresponding sections with a star * in the margin. Readers unfamiliar with these techniques may skip these sections and problems without loss of continuity. However, those with the necessary knowledge can gain more insight by reading the optional starred sections.

A set of problems is given at the end of each chapter and at the end of the book. The balance between theory and practice is evident in the list of problems, which vary in difficulty and purpose. These problems cover many areas of application and test understanding, theory, application, interpretation and the ability to read publications that use GLMs.

This book begins with an introduction to multiple linear regression. In a book about GLMs, at least three reasons exist for beginning with a short discussion of multiple linear regression:

- Linear regression is *familiar*. Starting with regression consolidates this material and establishes common notation, terminology and knowledge for all readers. Notation and new terms are best introduced in a familiar context.
- Linear regression is *foundational*. Many concepts and ideas from linear regression are used as approximations in GLMs. A firm foundation in linear regression ensures a better understanding of GLMs.
- Linear regression is *motivational*. GLMs often *improve* linear regression. Studying linear regression reveals its weaknesses and shows how GLMs can often overcome most of these, motivating the need for GLMs.

Connections between linear regression and GLMs are emphasized throughout this book.

This book contains a number of important but advanced topics and tools that have not typically been included in introductions to GLMs before. These include Tweedie family distributions with power variance functions, saddlepoint approximations, likelihood score tests, modified profile likelihood and randomized quantile residuals, as well as regression splines and orthogonal polynomials. Particular features are the use of saddlepoint approximations to clarify the asymptotical distribution of residual deviances from GLMs and an explanation of the relationship between score tests and Pearson statistics. Practical and specific guidelines are developed for the use of asymptotic approximations.

Throughout this book, R functions are shown in `typewriter font` followed by parentheses; for example, `glm()`. Operators, data frames and variables in R are shown in `typewriter font`; for example, `Smoke`. R packages are shown in **bold and sans serif font**; for example, **GLMsData**.

We thank those who have contributed to the writing of this book and especially students who have contributed to earlier versions of this text. We particularly thank Janette Benson, Alison Howes and Martine Maron for the permission to use data.

This book was prepared using LaTeX and R version 3.4.3 [3], integrated using Sweave [2].

Sippy Downs, QLD, Australia
Parkville, VIC, Australia
December 2017

Peter K. Dunn
Gordon K. Smyth

References

[1] Dunn, P.K., Smyth, G.K.: GLMsData: Generalized linear model data sets (2017). URL https://CRAN.R-project.org/package=GLMsData. R package version 1.0.0

[2] Leisch, F.: Dynamic generation of statistical reports using literate data analysis. In: W. Härdle, B. Rönz (eds.) Compstat 2002—Proceedings in Computational Statistics, pp. 575–580. Physika Verlag, Heidelberg, Germany (2002)

[3] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). URL https://www.R-project.org

[4] Reese, R.A.: Data analysis: The need for models? The Statistician **35**(2), 199–206 (1986). Special Issue: Statistical Modelling

Contents

Chapter 1
Statistical Models

> *. . . all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.*
> *Box and Draper [2, p. 424]*

1.1 Introduction and Overview

This chapter introduces the concept of a statistical model. One particular type of statistical model—the generalized linear model—is the focus of this book, and so we begin with an introduction to statistical models in general. This allows us to introduce the necessary language, notation, and other important issues. We first discuss conventions for describing data mathematically (Sect. 1.2). We then highlight the importance of plotting data (Sect. 1.3), and explain how to numerically code non-numerical variables (Sect. 1.4) so that they can be used in mathematical models. We then introduce the two components of a statistical model used for understanding data (Sect. 1.5): the systematic and random components. The class of regression models is then introduced (Sect. 1.6), which includes all models in this book. Model interpretation is then considered (Sect. 1.7), followed by comparing physical models and statistical models (Sect. 1.8) to highlight the similarities and differences. The purpose of a statistical model is then given (Sect. 1.9), followed by a description of the two criteria for evaluating statistical models: accuracy and parsimony (Sect. 1.10). The importance of understanding the limitations of statistical models is then addressed (Sect. 1.11), including the differences between observational and experimental data. The generalizability of models is then discussed (Sect. 1.12). Finally, we make some introductory comments about using R for statistical modelling (Sect. 1.13).

1.2 Conventions for Describing Data

The concepts in this chapter are best introduced using an example.

Example 1.1. A study of 654 youths in East Boston [10, 18, 20] explored the relationships between lung capacity (measured by forced expiratory volume,

P. K. Dunn, G. K. Smyth, *Generalized Linear Models with Examples in R*,
Springer Texts in Statistics, https://doi.org/10.1007/978-1-4419-0118-7_1

or FEV, in litres) and smoking status, age, height and gender (Table 1.1). The data are available in R as the data frame `lungcap` (short for 'lung capacity'), part of the **GLMsData** package [4]. For information about this package, see Appendix B; for more information about R, see Appendix A. Assuming the **GLMsData** package is installed in R (see Sect. A.2.4), load the **GLMsData** package and the `lungcap` data frame as follows:

```
> library(GLMsData)     # Load the  GLMsData  package
> data(lungcap)         # Make the data set  lungcap  available for use
> head(lungcap)         # Show the first few lines of data
  Age   FEV Ht Gender Smoke
1   3 1.072 46      F     0
2   4 0.839 48      F     0
3   4 1.102 48      F     0
4   4 1.389 48      F     0
5   4 1.577 49      F     0
6   4 1.418 49      F     0
```

(The `#` character and all subsequent text is ignored by R.) The data frame `lungcap` consist of five variables: `Age`, `FEV`, `Ht`, `Gender` and `Smoke`. Some of these variables are numerical variables (such as `Age`), and some are non-numerical variables (such as `Gender`). Any one of these can be accessed individually using `$` as follows:

```
> head(lungcap$Age)     # Show first six values of  Age
[1] 3 4 4 4 4 4
> tail(lungcap$Gender)  # Show last six values of  Gender
[1] M M M M M M
Levels: F M
```

Table 1.1 The forced expiratory volume (FEV) of youths, sampled from East Boston during the middle to late 1970s. FEV is in L; age is in completed years; height is in inches. The complete data set consists of 654 observations in total (Example 1.1)

Non-smokers						Smokers					
Females			Males			Females			Males		
FEV	Age	Height	FEV	Age	Height	FEV	Age	Height	FEV	Age	Height
1.072	3	46.0	1.404	3	51.5	2.975	10	63.0	1.953	9	58.0
0.839	4	48.0	0.796	4	47.0	3.038	10	65.0	3.498	10	68.0
1.102	4	48.0	1.004	4	48.0	2.387	10	66.0	1.694	11	60.0
1.389	4	48.0	1.789	4	52.0	3.413	10	66.0	3.339	11	68.5
1.577	4	49.0	1.472	5	50.0	3.120	11	61.0	4.637	11	72.0
1.418	4	49.0	2.115	5	50.0	3.169	11	62.5	2.304	12	66.5
1.569	4	50.0	1.359	5	50.5	3.102	11	64.0	3.343	12	68.0
1.196	5	46.5	1.776	5	51.0	3.069	11	65.0	3.751	12	72.0
1.400	5	49.0	1.452	5	51.0	2.953	11	67.0	4.756	13	68.0
1.282	5	49.0	1.930	5	51.0	3.104	11	67.5	4.789	13	69.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The length of any one variable is found using `length()`:

```
> length(lungcap$Age)
[1] 654
```

The dimension of the data set is:

```
> dim(lungcap)
[1] 654   5
```

That is, there are 654 cases and 5 variables. □

For these data, the sample size, usually denoted as n, is $n = 654$. Each youth's information is recorded in one row of the R data frame. FEV is called the *response variable* (or the *dependent variable*) since FEV is assumed to change in response to (or depends on) the values of the other variables. The response variable is usually denoted by y. In Example 1.1, y refers to 'FEV (in litres)'. When necessary, y_i refers to the ith value of the response. For example, $y_1 = 1.072$ in Table 1.1. Occasionally it is convenient to refer to all the observations y_i together instead of one at a time.

The other variables—age, height, gender and smoking status—can be called candidate variables, carriers, exogenous variables, independent variables, input variables, predictors, or regressors. We call these variables *explanatory variables* in this book. Explanatory variables are traditionally denoted by x. In Example 1.1, let x_1 refer to age (in completed years), and x_2 refer to height (in inches). When necessary, the value of, say, x_2 for Observation i is denoted x_{2i}; for example, $x_{2,1} = 46$.

Distinguishing between quantitative and qualitative explanatory variables is essential. Explanatory variables that are qualitative, like gender, are called *factors*. Gender is a factor with two *levels*: `F` (female) and `M` (male). Explanatory variables that are quantitative, like height and age, are called *covariates*.

Often, the key question of interest in an analysis concerns the relationship between the response variable and one or more explanatory variables, though other explanatory variables are present and may also influence the response. Adjusting for the effects of other correlated variables is often necessary, so as to understand the effect of the variable of key interest. These other variables are sometimes called *extraneous variables*. For example, we may be interested in the relationship between FEV (as the response variable) and smoking status (as the explanatory variable), but acknowledge that age, height and gender may also influence FEV. Age, height and gender are extraneous variables.

Example 1.2. Viewing the *structure* of a data frame can be informative:

```
> str(lungcap)          # Show the *structure* of the data frame
'data.frame':        654 obs. of  5 variables:
 $ Age   : int  3 4 4 4 4 4 4 5 5 5 ...
 $ FEV   : num  1.072 0.839 1.102 1.389 1.577 ...
 $ Ht    : num  46 48 48 48 49 49 50 46.5 49 49 ...
 $ Gender: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Smoke : int  0 0 0 0 0 0 0 0 0 0 ...
```

The size of the data frame is given, plus information about each variable: `Age` and `Smoke` consists of integers, `FEV` and `Ht` are numerical, while `Gender` is a factor with two *levels*. Each variable can be summarized numerically using `summary()`:

```
> summary(lungcap)     # Summarize the data
      Age              FEV              Ht          Gender
 Min.   : 3.000   Min.   :0.791   Min.   :46.00   F:318
 1st Qu.: 8.000   1st Qu.:1.981   1st Qu.:57.00   M:336
 Median :10.000   Median :2.547   Median :61.50
 Mean   : 9.931   Mean   :2.637   Mean   :61.14
 3rd Qu.:12.000   3rd Qu.:3.119   3rd Qu.:65.50
 Max.   :19.000   Max.   :5.793   Max.   :74.00
     Smoke
 Min.   :0.00000
 1st Qu.:0.00000
 Median :0.00000
 Mean   :0.09939
 3rd Qu.:0.00000
 Max.   :1.00000
```

Notice that quantitative variables are summarized differently to qualitative variables. `FEV`, `Age` and `Ht` (all quantitative) are summarized with the minimum and maximum values, the first and third quartiles, and the mean and median. `Gender` (qualitative) is summarised by giving the number of males and females in the data. The variable `Smoke` is qualitative, and numbers are used to designate the levels of the variable. In this case, R has no way of determining if the variable is a factor or not, and assumes the variable is quantitative by default since it consists of numbers. To explicitly tell R that `Smoke` is qualitative, use `factor()`:

```
> lungcap$Smoke <- factor(lungcap$Smoke,
                  levels=c(0, 1),                  # The values of  Smoke
                  labels=c("Non-smoker","Smoker")) # The labels
> summary(lungcap$Smoke)     # Now, summarize the redefined variable Smoke
Non-smoker     Smoker
       589         65
```

(The information about the data set, accessed using `?lungcap`, explains that `0` represents non-smokers and `1` represents smokers.) We notice that non-smokers outnumber smokers. □

1.3 Plotting Data

Understanding the lung capacity data is difficult because there is so much data. How can the impact of age, height, gender and smoking status on FEV be understood? Plots (Fig. 1.1) may reveal many, but probably not all, important features of the data:

```
> plot( FEV ~ Age, data=lungcap,
       xlab="Age (in years)",      # The x-axis label
       ylab="FEV (in L)",          # The y-axis label
       main="FEV vs age",          # The main title
       xlim=c(0, 20),              # Explicitly set x-axis limits
       ylim=c(0, 6),               # Explicitly set y-axis limits
       las=1)                      # Makes axis labels horizontal
```

This R code uses the `plot()` command to produce plots of the data. (For more information on plotting in R, see Sect. A.3.10.) The formula `FEV ~ Age` is read as '`FEV` is modelled by `Age`'. The input `data=lungcap` indicates that `lungcap` is the data frame in which to find the variables `FEV` and `Age`. Continue by plotting `FEV` against the remaining variables:

```
> plot( FEV ~ Ht, data=lungcap, main="FEV vs height",
       xlab="Height (in inches)", ylab="FEV (in L)",
       las=1, ylim=c(0, 6) )
> plot( FEV ~ Gender, data=lungcap,
       main="FEV vs gender", ylab="FEV (in L)",
       las=1, ylim=c(0, 6))
> plot( FEV ~ Smoke,  data=lungcap, main="FEV vs Smoking status",
       ylab="FEV (in L)", xlab="Smoking status",
       las=1, ylim=c(0, 6))
```

(Recall that `Smoke` was declared a factor in Example 1.2.) Notice that R uses different types of displays for plotting FEV against covariates (top panels) than against factors (bottom panels). Boxplots are used (by default) for plotting FEV against factors: the solid horizontal centre line in each box represents the median (not the mean), and the limits of the central box represent the upper and lower quartiles of the data (approximately 75% of the observations are less than the upper quartile, and approximately 25% of the observations are less than the lower quartile). The lines from the central box extend to the largest and smallest values, except for outliers which are indicated by individual points (such as a large FEV for a few smokers). In R, outliers are defined, by default, as observations more than 1.5 times the interquartile range (the difference between the upper and lower quartiles) more extreme than the upper or lower limits of the central box.

The plots (Fig. 1.1) show a moderate relationship (reasonably large variation) between FEV and age, that is possibly linear (at least until about 15 years of age). However, a stronger relationship (less variation) is apparent between FEV and height, but this relationship does not appear to be linear.

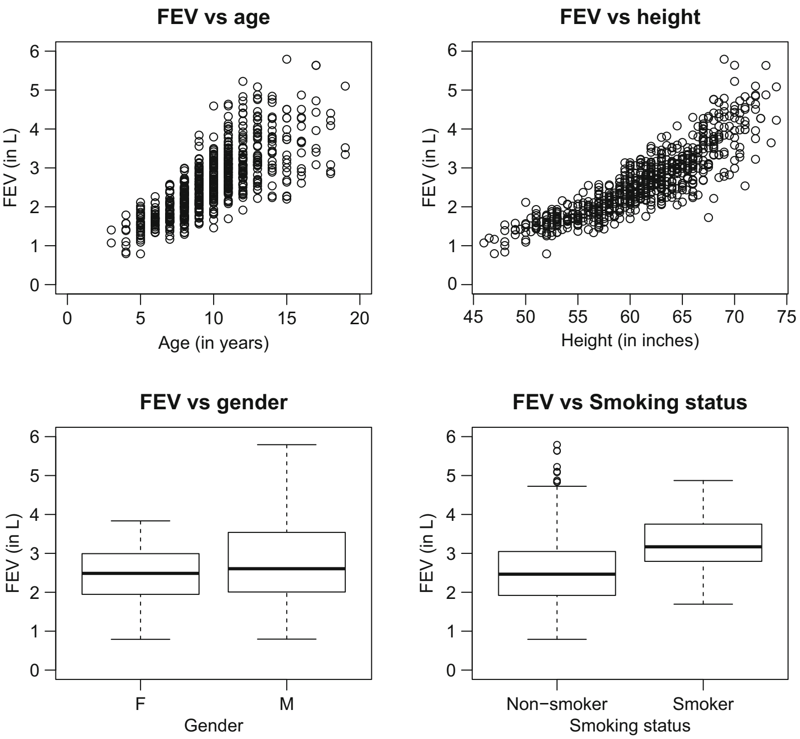

Fig. 1.1 Forced expiratory volume (FEV) plotted against age (top left), height (top right), gender (bottom left) and smoking status (bottom right) for the data in Table 1.1 (Sect. 1.3)

The variation in FEV appears to increase for larger values of FEV also. In general, it also appears that males have a slightly larger FEV, and show greater variation in FEV, than females. Smokers appear to have a larger FEV than non-smokers.

While many of these statements are expected, the final statement is surprising, and may suggest that more than one variable should be examined at once. The plots in Fig. 1.1 only explore the relationships between FEV and each explanatory variable individually, so we continue by exploring relationships involving more than two variables at a time.

One way to do this is to plot the data separately for smokers and non-smokers (Fig. 1.2), using similar scales on the axes to enable comparisons:

```
> plot( FEV ~ Age,
    data=subset(lungcap, Smoke=="Smoker"),  # Only select smokers
    main="FEV vs age\nfor smokers",         # \n means `new line'
    ylab="FEV (in L)", xlab="Age (in years)",
    ylim=c(0, 6), xlim=c(0, 20), las=1)
```

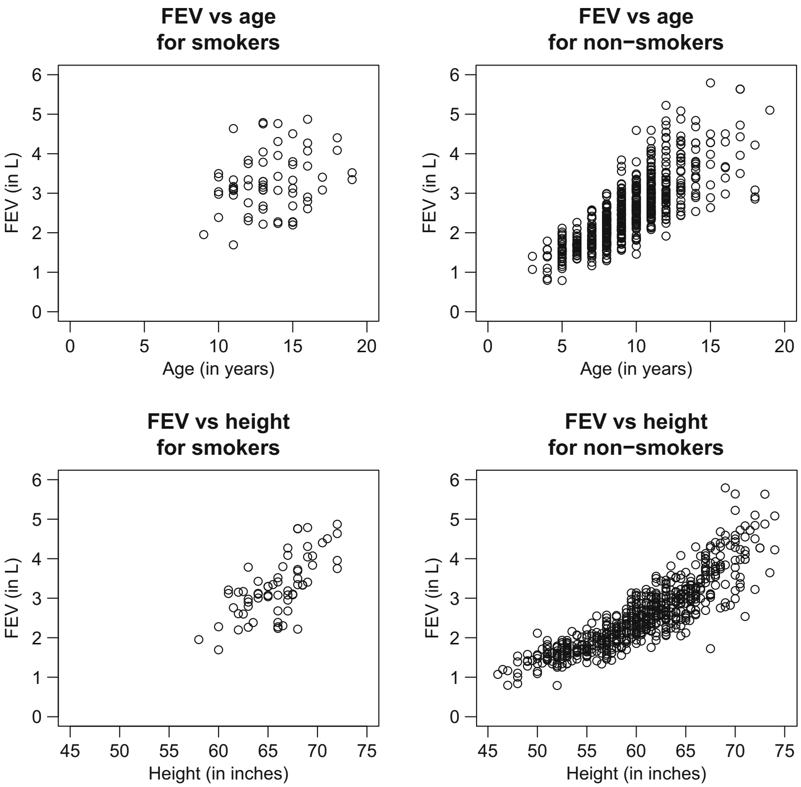

Fig. 1.2 Plots of the lung capacity data: the forced expiratory volume (FEV) plotted against age, for smokers (top left panel) and non-smokers (top right panel); and the forced expiratory volume (FEV) plotted against height, for smokers (bottom left panel) and non-smokers (bottom right panel) (Sect. 1.3)

```
> plot( FEV ~ Age,
    data=subset(lungcap, Smoke=="Non-smoker"),  # Only select non-smokers
    main="FEV vs age\nfor non-smokers",
    ylab="FEV (in L)", xlab="Age (in years)",
    ylim=c(0, 6), xlim=c(0, 20), las=1)
> plot( FEV ~ Ht, data=subset(lungcap, Smoke=="Smoker"),
    main="FEV vs height\nfor smokers",
    ylab="FEV (in L)", xlab="Height (in inches)",
    xlim=c(45, 75), ylim=c(0, 6), las=1)
> plot( FEV ~ Ht, data=subset(lungcap, Smoke=="Non-smoker"),
    main="FEV vs height\nfor non-smokers",
    ylab="FEV (in L)",  xlab="Height (in inches)",
    xlim=c(45, 75), ylim=c(0, 6), las=1)
```

Note that == is used to make logical comparisons. The plots show that smokers tend to be older (and hence taller) than non-smokers and hence are likely to have a larger FEV.

Another option is to distinguish between smokers and non-smokers when plotting the FEV against Age. For these data, there are so many observations that distinguishing between smokers and non-smokers is difficult, so we first adjust Age so that the values for smokers and non-smokers are slightly separated:

```
> AgeAdjust <- lungcap$Age + ifelse(lungcap$Smoke=="Smoker", 0, 0.5)
```

The code `ifelse( lungcap$Smoke=="Smoker", 0, 0.5)` adds zero to the value of Age for youth labelled with Smoker, and adds 0.5 to youth labelled otherwise (that is, non-smokers). Then we plot FEV against this variable: (Fig. 1.3, top left panel):

```
> plot( FEV ~ AgeAdjust, data=lungcap,
    pch = ifelse(Smoke=="Smoker", 3, 20),
    xlab="Age (in years)", ylab="FEV (in L)", main="FEV vs age", las=1)
```

The input pch indicates the plotting character to use when plotting; then, `ifelse( Smoke=="Smoker", 3, 20)` means to plot with plotting character 3 (a 'plus' sign) if Smoke takes the value "Smoker", and otherwise to plot with plotting character 20 (a filled circle). See ?points for an explanation of the numerical codes used to define different plotting symbols. Recall that in Example 1.2, Smoke was declared as a factor with two levels that were labelled Smoker and Non-smoker. The legend() command produces the legend:

```
> legend("topleft", pch=c(20, 3), legend=c("Non-smokers","Smokers") )
```

The first input specifies the location (such as "center" or "bottomright"). The second input gives the plotting notation to be explained (such as the points, using pch, or the line types, using lty). The legend input provides the explanatory text. Use ?legend for more information.

A boxplot can also be used to show relationships (Fig. 1.3, top right panel):

```
> boxplot(lungcap$FEV ~ lungcap$Smoke + lungcap$Gender,
    ylab="FEV (in L)", main="FEV, by gender\n and smoking status",
    las=2,    # Keeps labels perpendicular to the axes
    names=c("F:\nNon", "F:\nSmoker", "M:\nNon", "M:\nSmoker"))
```

Another way to show the relationship between three variables is to use an *interaction plot*, which shows the relationship between the levels of two factors and (by default) the mean response of a quantitative variable. The appropriate R function is interaction.plot() (Fig. 1.3, bottom panels):

```
> interaction.plot( lungcap$Smoke, lungcap$Gender, lungcap$FEV,
                    xlab="Smoking status", ylab="FEV (in L)",
                    main="Mean FEV, by gender\n and smoking status",
                    trace.label="Gender", las=1)
> interaction.plot( lungcap$Smoke, lungcap$Gender, lungcap$Age,
                    xlab="Smoking status", ylab="Age (in years)",
                    main="Mean age, by gender\n and smoking status",
                    trace.label="Gender", las=1)
```

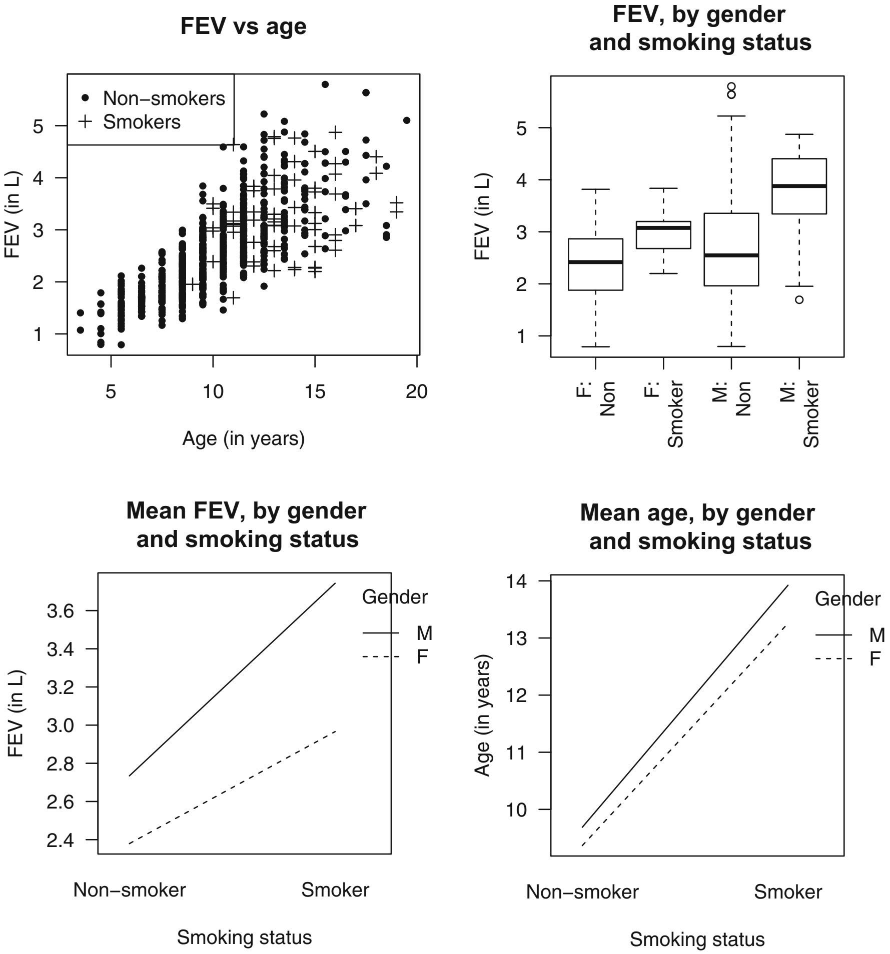

Fig. 1.3 Plots of the lung capacity data: the forced expiratory volume (FEV) plotted against age, using different plotting symbols for non-smokers and smokers (top left panel); a boxplot of FEV against gender and smoking status (top right panel); an interaction plot of the mean FEV against smoking status according to gender (bottom left panel); and an interaction plot of the mean age against smoking status according to gender (bottom right panel) (Sect. 1.3)

This plot shows that, in general, smokers have a larger FEV than non-smokers, for both males and females. The plot also shows that the mean age of smokers is higher for both males and females.

To make any further progress quantifying the relationship between the variables, mathematics is necessary to create a *statistical model*.

1.4 Coding for Factors

Factors represent categories (such as smokers or non-smokers, or males and females), and so must be coded numerically to be used in mathematical models. This is achieved by using *dummy variables*.

The variable `Gender` in the `lungcap` data frame is loaded as a factor by default, as the data are non-numerical:

```
> head(lungcap$Gender)
[1] F F F F F F
Levels: F M
```

To show the coding used by R for the variable `Gender` in the `lungcap` data set, use `contrasts()`:

```
> contrasts(lungcap$Gender)
  M
F 0
M 1
```

(The function name is because, under certain conditions, the codings are called contrasts.) The output shows the two levels of `Gender` on the left, and the name of the dummy variable across the top. When the dummy variable `M` is equal to one, the dummy variable refers males. Notice `F` is not listed across the top of the output as a dummy variable, since it is the *reference level*. By default in R, the reference level is the first level alphabetically or numerically. In other words, the dummy variable, say x_3, is:

$$x_3 = \begin{cases} 0 & \text{if } \texttt{Gender} \text{ is } \texttt{F} \text{ (females)} \\ 1 & \text{if } \texttt{Gender} \text{ is } \texttt{M} \text{ (males).} \end{cases} \tag{1.1}$$

Since these numerical codes are arbitrarily assigned, other levels may be set as the reference level in R using `relevel()`:

```
> contrasts( relevel( lungcap$Gender, "M") )  # Now, M is the ref. level
  F
M 0
F 1
```

As seen earlier in Example 1.2, the R function `factor()` is used to explicitly declare a variable as a factor when necessary (for example, if the data use numbers to designate the factor levels):

```
> lungcap$Smoke <- factor(lungcap$Smoke,
                          levels=c(0, 1),
                          labels=c("Non-smoker","Smoker"))
> contrasts(lungcap$Smoke)
           Smoker
Non-smoker      0
Smoker          1
```

This command assigns the values of `0` and `1` to the labels `Non-smoker` and `Smoker` respectively:

$$x_4 = \begin{cases} 0 & \text{if } \texttt{Smoke} \text{ is } \texttt{0} \text{ (non-smoker)} \\ 1 & \text{if } \texttt{Smoke} \text{ is } \texttt{1} \text{ (smokers).} \end{cases} \tag{1.2}$$

For a factor with k levels, $k-1$ dummy variables are needed. For example, if smoking status had three levels (for example, 'Never smoked', 'Former smoker', 'Current smoker'), then two dummy variables are needed:

$$x_5 = \begin{cases} 1 & \text{for former smokers} \\ 0 & \text{otherwise;} \end{cases} \qquad x_6 = \begin{cases} 1 & \text{for current smokers} \\ 0 & \text{otherwise.} \end{cases} \tag{1.3}$$

Then $x_5 = x_6 = 0$ uniquely refers to people who have never smoked.

The coding discussed here is called *treatment coding*. Many types of coding exist to numerically code factors. Treatment coding is commonly used (and is used in this book, and in R by default) since it usually leads to a direct interpretation. Other codings are also possible, with different interpretations useful in different contexts. In any analysis, the definition of the dummy variables being used should be made clear.

1.5 Statistical Models Describe Both Random and Systematic Features of Data

Consider again the lung capacity data from Example 1.1 (p. 1). At any given combination of height, age, gender and smoking status, many different values of FEV could be recorded, and so produce a *distribution* of recorded FEV values. A model for this distribution of values is called the *random component* of the statistical model. At this given combination of height, age, gender and smoking status, the distribution of FEV values has a mean FEV. The mathematical relationship between the mean FEV and given values of height, age, gender and smoking status is called the *systematic component* of the model. A statistical model consists of a random component and a systematic component to explain these two features of real data. In this context, the *role* of a statistical model is to mathematically represent both the systematic and random components of data.

Many systematic components for the lung capacity data are possible. One simple systematic component is

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \tag{1.4}$$

for Observation i, where μ_i is the *expected value* of y_i, so that $\mu_i = \mathrm{E}[y_i]$ for $i = 1, 2, \ldots, n$. The β_j (for $j = 0, 1, 2, 3$ and 4) are unknown *regression parameters*. The explanatory variables are age x_1, height x_2, the dummy variable x_3 defined in (1.1) for gender, and the dummy variable x_4 defined in (1.2) for smoking status. This is likely to be a poor systematic component, as the plots (Fig. 1.1) show that the relationship between FEV and height is non-linear, for example. Other systematic components are also possible.

The randomness about this systematic component may take many forms. For example, using $\text{var}[y_i] = \sigma^2$ assumes that the variance of the responses y_i is constant about μ_i, but makes no assumptions about the distribution of the responses. A popular assumption is to assume the responses have a normal distribution about the mean μ_i with constant variance σ^2, written $y_i \sim N(\mu_i, \sigma^2)$, where '$\sim$' means 'is distributed as'. Both assumptions are likely to be poor for the lung capacity data, as the plots (Fig. 1.1) show that the variation in the observed FEV increases for larger values of FEV. Other assumptions are also possible, such as assuming the responses come from other probability distributions beside the normal distribution.

1.6 Regression Models

The systematic component (1.4) for the lung capacity data is one possible representation for explaining how the mean FEV changes as height, age, gender and smoking status vary. Many other representation are also possible. Very generally, a *regression model* assumes that the mean response μ_i for Observation i depends on the p explanatory variables x_{1i} to x_{pi} via some general function f through a number of regression parameters β_j (for $j = 0, 1, \dots q$). Mathematically,

$$\text{E}[y_i] = \mu_i = f(x_{1i}, \dots, x_{pi}; \beta_0, \beta_1, \dots, \beta_q).$$

Commonly, the parameters β_j are assumed to combine the effects of the explanatory variables linearly, so that the systematic component often takes the more specific form

$$\mu_i = f(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}). \tag{1.5}$$

Regression models with this form (1.5) are *regression models linear in the parameters*. All the models discussed in this book are regression models linear in the parameters. The component $\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$ is called the *linear predictor*.

Two special types of regression models linear in the parameters are discussed in detail in this book:

- Linear regression models: The systematic component of a linear regression model assumes the form

$$\text{E}[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \tag{1.6}$$

while the randomness is assumed to have constant variance σ^2 about μ_i. Linear regression models are formally defined and discussed in Chaps. 2 and 3.

- Generalized linear models: The systematic component of a generalized linear model assumes the form

$$\begin{aligned} \mu_i &= g^{-1}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) \\ \text{or alternatively: } g(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \end{aligned}$$

  where $g()$ (called a *link function*) is a monotonic, differentiable function (such as a logarithm function). The randomness is explained by assuming y has a distribution from a specific family of probability distributions (which includes common distributions such as the normal, Poisson and binomial as special cases). Generalized linear models are discussed from Chap. 5 onwards. An example of a generalized linear model appears in Example 1.5. Linear regression models are a special case of generalized linear models.

The following notational conventions apply to regression models linear in the parameters:

- The number of explanatory variables is p: x_1, x_2, $\ldots x_p$.
- The number of regression parameters is denoted p'. If a constant term β_0 is in the systematic component (as is almost always the case) then $p' = p+1$, and the regression parameters are β_0, β_1, $\ldots \beta_p$. If a constant term β_0 is *not* in the systematic component then $p' = p$, and the regression parameters are β_1, β_2, $\ldots \beta_p$.

Example 1.3. For the lungcap data (Example 1.1, p. 1), a possible systematic component is given in (1.4) for some numerical values of β_0, β_1, β_2, β_3 and β_4, for $i = 1, 2, \ldots, 654$. This systematic relationship implies a *linear* relationship between μ and the covariates Age x_1 (which may be reasonable from Fig. 1.1, top left panel), and Height x_2, (which is probably *not* reasonable from Fig. 1.1, top right panel). The model has $p = 4$ explanatory variables, and $p' = 5$ unknown regression parameters.

One model for the random component, suggested in Sect. 1.5, was that the variation of the observations about this systematic component was assumed to be approximately constant, so that $\operatorname{var}[y_i] = \sigma^2$. Combining the two components, a possible linear regression model for modelling the FEV is

$$\begin{cases} \operatorname{var}[y_i] = \sigma^2 & \text{(random component)} \\ \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} & \text{(systematic component).} \end{cases} \tag{1.7}$$

Often the subscripts i are dropped for simplicity when there is no ambiguity. The values of the parameters β_0, β_1, β_2, β_3, β_4 (for the systematic component) and σ^2 (for the random component) are unknown, and must be estimated.

This is the model implied in Sect. 1.5, where it was noted that both the systematic and random components in (1.7) are likely to be inappropriate for these data (Fig. 1.1). □

Example 1.4. Some other possible systematic components involving FEV (y), age (x_1), height (x_2), gender (x_3) and smoking status (x_4) include:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 \tag{1.8}$$

$$\mu = \beta_0 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_4 \tag{1.9}$$

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{1.10}$$

$$\mu = \beta_0 + \beta_1 \log x_1 + \beta_2 x_2 + \beta_4 x_4 \tag{1.11}$$

$$\mu = \beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_4 \tag{1.12}$$

$$1/\mu = \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 \tag{1.13}$$

$$\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 \tag{1.14}$$

$$\mu = \beta_0 + \exp(\beta_1 x_1) - \exp(\beta_2 x_2) + \beta_4 x_4^2 \tag{1.15}$$

All these systematic components apart from (1.15) are linear in the parameters and could be used as the systematic component of a generalized linear model. Only (1.8)–(1.12) could be used to specify a linear regression model. □

Example 1.5. The noisy miner is a small but aggressive native Australian bird. A study [11] of the habitats of the noisy miner recorded (Table 1.2; data set: `nminer`) the abundance of noisy miners (that is, the number observed; `Minerab`) in two hectare transects located in buloke woodland patches with varying numbers of eucalypt trees (`Eucs`). To plot the data (Fig. 1.4), a small amount of randomness is first added in the vertical direction to avoid over plotting, using `jitter()`:

```
> data(nminer)      # Load the data
> names(nminer)     # Show the variables
[1] "Miners"  "Eucs"    "Area"    "Grazed"  "Shrubs"  "Bulokes" "Timber"
[8] "Minerab"
> plot( jitter(Minerab) ~ Eucs, data=nminer, las=1, ylim=c(0, 20),
   xlab="Number of eucalypts per 2 ha", ylab="Number of noisy miners" )
```

See `?nminer` for more information about the data and the other variables.

The random component certainly does not have constant variance, as the observations are more spread out for a larger numbers of eucalypts. Because the responses are counts, a *Poisson distribution* with mean μ_i for Observation i may be suitable for modelling the data. We write $y_i \sim \text{Pois}(\mu_i)$, where $\mu_i > 0$.

The relationship between μ and the number of eucalypts also seems nonlinear. A possible model for the systematic component is $\text{E}[y_i] = \mu_i = \exp(\beta_0 + \beta_1 x_i)$, where x_i is the number of eucalypt trees at location i. This

Table 1.2 The number of eucalypt trees and the number of noisy miners observed in two hectare transects in buloke woodland patches within the Wimmera Plains of western Victoria, Australia (Example 1.5)

Number of eucalypts	Number of noisy miners	Number of eucalypts	Number of noisy miners	Number of eucalypts	Number of noisy miners
2	0	32	19	0	0
10	0	2	0	0	0
16	3	16	2	0	0
20	2	7	0	3	0
19	8	10	3	8	0
18	1	15	1	8	0
12	8	30	7	15	0
16	5	4	1	21	3
3	0	4	0	24	4
12	4	19	7	15	6
		11	0		

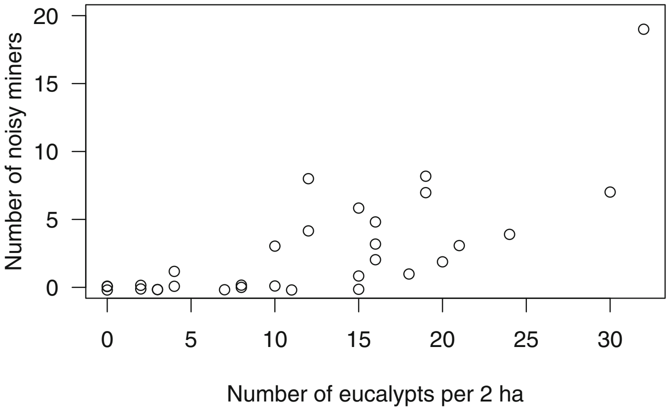

Fig. 1.4 The number of noisy miners (observed in two hectare transects in buloke woodland patches within the Wimmera Plains of western Victoria, Australia) plotted against the number of eucalypt trees. A small amount of randomness is added to the number of miners in the vertical direction to avoid over-plotted observations (Example 1.5)

functional form ensures $\mu_i > 0$, as required for the Poisson distribution, and may also be appropriate for modelling the non-linearity.

Combining the two components, one possible model for the data, dropping the subscripts i, is:

$$\begin{cases} y \sim \text{Pois}(\mu) & \text{(random component)} \\ \mu = \exp(\beta_0 + \beta_1 x) & \text{(systematic component)} \end{cases} \tag{1.16}$$

where $\mu = \text{E}[y]$. This is an example of a *Poisson generalized linear model* (Chap. 10).

We also note that one location (with 19 noisy miners) has more than twice the number of noisy miners observed than the location with the next largest number of noisy miners (with eight noisy miners). □

1.7 Interpreting Regression Models

Models are most useful when they have sensible interpretations. Compare these two systematic components:

$$\mu = \beta_0 + \beta_1 x \tag{1.17}$$

$$\log \mu = \beta_0 + \beta_1 x. \tag{1.18}$$

The first model (1.17) assumes a linear relationship between μ and x, and hence that an increase of one in the value of x is associated with an increase of β_1 in the value of μ. The second model (1.18) assumes a linear relationship between $\log \mu$ and x, and hence that an increase of one in the value of x will increase the value of $\log \mu$ by β_1. This implies that when the value of x increases by one, μ increases (approximately) by a *factor* of $\exp(\beta_1)$. To see this, write the second systematic component (1.18) as

$$\mu_x = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0)\exp(\beta_1)^x.$$

Hence if the value of x increases by 1, to $x + 1$, we have

$$\mu_{x+1} = \exp(\beta_0)\exp(\beta_1)^{x+1} = \mu_x \exp(\beta_1).$$

A researcher should consider which is more sensible for the application. Furthermore, models that are based on underlying theory or sensible approximations to the problem (Sect. 1.10) produce models with better and more meaningful interpretations. Note that the systematic component (1.17) is suitable for a linear regression model, and that both systematic components are suitable for a generalized linear model.

Example 1.6. For the `lungcap` data, consider a model relating FEV y to height x. Model (1.17) would imply that an increase in height of one inch is associated with an increase in FEV of β_1 L. In contrast, Model (1.18) would imply that an increase in height of one inch is associated with an increase in FEV by a factor of $\exp(\beta_1)$ L. □

A further consideration when interpreting models is when models contain more than one explanatory variable. In these situations, the regression parameters should be interpreted with care, since the explanatory variables may not be independent. For example, for the lung capacity data, the age and height of youth are related (Fig. 1.5): older youth are taller, on average:

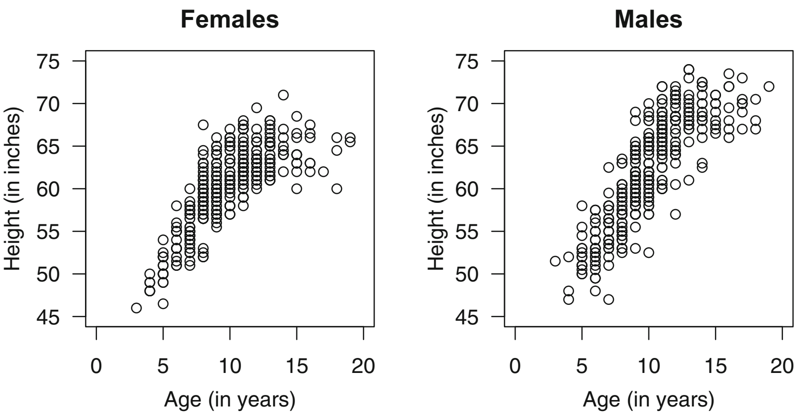

Fig. 1.5 A strong relationship exists between the height and the age of the youth in the lung capacity data: females (left panel) and males (right panel)

```
> plot( Ht ~ Age, data=subset(lungcap, Gender=="F"), las=1,
    ylim=c(45, 75), xlim=c(0, 20), # Use similar scales for comparisons
    main="Females", xlab="Age (in years)", ylab="Height (in inches)" )
> plot( Ht ~ Age, data = subset(lungcap, Gender=="M"), las=1,
    ylim=c(45, 75), xlim=c(0, 20), # Use similar scales for comparisons
    main="Males", xlab="Age (in years)", ylab="Height (in inches)" )
```

In a model containing both age and height, it is not possible to interpret both regression parameters independently, as expecting age to change while height stays constant is unreasonable in youth. Note that height tends to increase with age initially, then tends to stay similar as the youth stop (or slow) their growing.

Further comments on model interpretation for specific models are given as appropriate, such as in Sect. 2.7.

1.8 All Models Are Wrong, but Some Are Useful

Previous sections introduced regression models as a way to understand data. However, when writing about statistical models, Box and Draper [2, p. 424] declared “all models are wrong”. What do they mean? Were they correct? One way to understand this is to contrast statistical models with some physical models in common use. For example, biologists use *models* of the human skeleton to teach anatomy, which capture enough important information about the real situation for the necessary purpose. Models are not an exact representation of reality: the skeleton is probably made of plastic, not bones; no-one may have a skeleton with the exact dimensions of the model skeleton. However, models *are* useful approximations for representing the necessary detail for the purpose at hand.

Similar principles apply to *statistical models*: they are mathematical approximations to reality that represent the important features of *data* for the task at hand. The complete quote from Box and Draper clarifies [2, p. 424], "...Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind".

Despite the many similarities between physical and statistical models, two important differences exist:

- A model skeleton shows the structure of an average or typical skeleton, which is equivalent to the systematic component of a statistical model. But no-one has a skeleton exactly like the model: some bones will be longer, skinnier, or a different shape. However, the model skeleton makes no attempt to indicate the *variation* that is present in skeletons in the population. The model skeleton ignores the variation from person to person (the random component). In contrast, the statistical model represents both the systematic trend and the randomness of the data. The random component is modelled explicitly by making precise statements about the random variation (Sect. 1.5).
- Most physical models are based on what is known to be true. Biologists *know* what a typical real skeleton looks like. Consequently, knowing whether a physical model is adequate is generally easy, since the model represents the important, known features of the true situation. However, statistical models are often developed where the *true* model is unknown, or is only artificially assumed to exist. In these cases, the model must be developed from the available data.

1.9 The Purpose of a Statistical Model Affects How It Is Developed: Prediction vs Interpretation

The *role* of a statistical model is to accurately represent the important systematic and random features of the data. But what is the *purpose* of developing statistical models? For regression models, there are two major motivations:

- Prediction: To produce accurate predictions from new or future data.
- Understanding and interpretation: To understand how variables relate to each other.

For example, consider the lung capacity study. The purpose of this study may be to determine whether there is a (potentially causal) relationship between smoking and FEV. Here we want to understand whether smoking has an effect on FEV, and in what direction. For this purpose, the size and significance of coefficients in the model are of interest. If smoking decreases lung function, this would have implications for health policy.

A different health application is to establish the normal weight range for children of a given age and gender. Here the purpose is to be able to judge whether a particular child is out of the normal range, in which case some intervention by health carers might be appropriate. In this case, a prediction curve relating weight to age is desired, but the particular terms in the model would not be of interest. The lung capacity data is in fact an extract from a larger study [19] in which the pulmonary function of the same children was measured at multiple time points (a *longitudinal study*), with the aim of establishing the normal range for FEV at each age.

Being aware of the major purpose of a study may affect how a regression model is fitted and developed. If the major purpose is interpretation, then it is important that all terms are reliably estimated and have good support from the data. If the major purpose is prediction, then any predictor that improves the precision of prediction may be included in the model, even if the causal relationship between the predictor and the response is obscure or if the regression coefficient is relatively uncertain. This means that sometimes one might include more terms in a regression model when the purpose is prediction than when the purpose is interpretation and understanding.

1.10 Accuracy vs Parsimony

For any set of data, there are typically numerous systematic components that could be chosen and various random components may also be possible. How do we choose a statistical model from all the possible options?

Sometimes, statistical models are based on underlying theory, or from an understanding of the physical features of the situation, and are built with this knowledge in mind. In these situations, the statistical model may be critiqued by how well the model explains the known features of the theoretical situation.

Sometimes, approximations to the problem can guide the choice of model. For example, for the lung capacity data, consider lungs roughly as cylinders, whose heights are proportional to the height of the child, and assume the FEV is proportional to lung volume. Then volume $\propto$ (radius)$^2 x_2$ may be a suitable model. This approach implies FEV is proportional to x_2, as in Models (1.8)–(1.11) (p. 14).

Sometimes, statistical models are based on data, often without guiding theory, and no known ‘true’ state exists with which to compare. After all, statistical models are artificial, mathematical constructs. The model is a representation of an unknown, but assumed, underlying true state. How can we know if the statistical model is adequate?

In general, an adequate statistical model balances two criteria:

- Accuracy: The model should accurately describe both the systematic and random components.
- Parsimony: The model should be as simple as possible.

According to the *principle of parsimony* (or *Occam's Razor*), the simplest accurate model is the preferred model. In other words, prefer the simplest accurate model not contradicting the data. A model too simple or too complex does not model the data well. Complex models may fit the given data well but usually do not generalize well to other data sets (this is called *over-fitting*).

Example 1.7. Figure 1.6 (top left panel) shows the systematic component of a linear model (represented by the solid line) fitted to some data. This model does not represent the systematic trend of the data. The variation around this linear model is large and not random: observations are consistently smaller than the fitted model, then consistently larger, then smaller.

The systematic component of the fitted cubic model (Fig. 1.6, top centre panel) represents the systematic trend of the data, and suggests a small amount of random variation about this trend.

The fitted 10th order polynomial (Fig. 1.6, top right panel) suggests a small amount of randomness, as the polynomial passes close to every observation. However, the systematic polynomial component incorrectly represents both the systematic *and* random components in the data. Because the systematic component also represents the randomness, predictions based on this model are suspect (predictions near $x = -1$ are highly dubious, for example).

The principle of parsimony suggests the cubic model is preferred. This model is simple, accurate, and does not contradict the data. Researchers focused only on producing a model passing close to each observation (and hence selecting the 10th order polynomial) have a poor model. This is called *over-fitting*.

The data were actually generated from the model

$$\begin{cases} y \sim N(\mu, 0.35) \\ \mu = x^3 - 3x + 5. \end{cases}$$

The notation $y \sim N(\mu, 0.35)$ means the responses come from a normal distribution with mean μ and variance $\sigma^2 = 0.35$.

Suppose new data were observed from this same true model (for example, from a new experiment or from a new sample), and linear, cubic and 10th order polynomial models were refitted to this new data (Fig. 1.6, bottom panels). The new fitted linear model (Fig. 1.6, bottom left panel) still does not fit the data well. The new fitted 10th order polynomial (Fig. 1.6, bottom right panel) is very different compared to the one fitted to the first data set, even though the data for both were generated from the same model. In contrast, the new fitted cubic model (Fig. 1.6, bottom centre panel) is very similar for both data sets, suggesting the cubic model represents the systematic and random components well. □

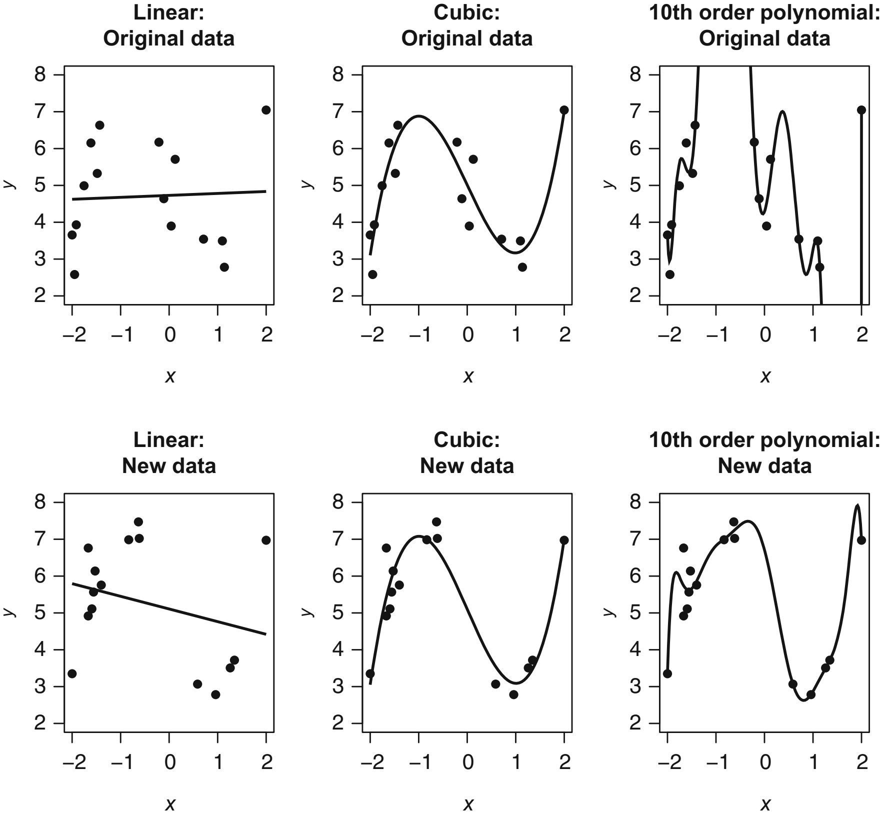

Fig. 1.6 Three different systematic components for an artificial data set. Left panels: the data modelled using a linear model; centre panels: using a cubic model; right panels: using a 10th order polynomial. The lines represent the systematic component of the fitted model. The top panels show the models fitted to some data; the bottom panels shows the models fitted to data randomly generated from the same model used to generate the data in the top panels. A good model would be similar for both sets of data (Example 1.7)

1.11 Experiments vs Observational Studies: Causality vs Association

All models must be used and understood within limitations imposed by how the data were collected. The method of data collection influences the conclusions that can be drawn from the analysis. An important aspect of this concerns whether researchers intervene to apply treatments to subjects or simply observe pre-existing processes.

In an *observational study*, researchers may use elaborate equipment to collect physical measures or may ask subjects to respond to carefully designed questionnaires, but do not influence the processes being observed.

Observational studies generally only permit conclusions about *associations* between variables, not a cause-and-effect. While the relationship may in fact be causal, the use of observational data by itself it not usually sufficient to confirm this conclusion. In contrast, researchers conducting a *designed experiment* do intervene to control the values of the explanatory variables that appear in the data. The distinguishing feature of an experiment versus an observational study is that the researchers conducting the study are able to determine which experimental condition is applied to each subject. A well-designed randomized experiment allows inference to be made about *cause-and-effect* relationships between the explanatory and response variables.

Statistical models treat experimental and observational studies in the same way, and the statistical conclusions are superficially similar, but scientific conclusions from experiments are usually much stronger. In an observational study, the best that can be done is to measure all other extraneous variables that are likely to affect the response, so that the analysis can adjust for as many uncontrolled effects as possible. In this way, good quality data and careful statistical analysis can go a long way towards correcting for many influences that cannot be controlled in the study design.

Example 1.8. The lung capacity data (Example 1.1) is a typical observational study. The purpose of the study is to explore the effects of smoking on lung capacity, as measured by FEV (explored later in Problem 11.15). Whether or not each participant is a smoker is out of the control of the study designers, and there are many physical characteristics, such as age and height, that have direct effects on lung capacity, and some quite probably have larger effects than the effect of interest (that of smoking). Hence it was necessary to record information on the height, age and gender of participants (which become extraneous variables) so that the influence of these variables can be taken into account. The aim of the analysis therefore is to try to measure the association between smoking and lung capacity after adjusting for age, height and gender. It is always possible that there are other important variables that influence FEV that have not been measured, so any association discovered between FEV and smoking should not be assumed to be cause-and-effect. □

1.12 Data Collection and Generalizability

Another feature of data collection that affects conclusions is the population from which the subjects or cases are drawn. In general, conclusions from fitting and analysing a statistical model only apply to the population from which the cases are drawn. So, for example, if subjects are drawn from women aged over 60 in Japan, then conclusions do not necessarily apply to men, to women in Japan aged under 60, or to women aged over 60 elsewhere.

Similarly, the conclusions from a regression model cannot necessarily be applied (extrapolated) outside the range of the data used to build the model.

Example 1.9. The lung capacity data (Example 1.1) is from a sample of youths from the middle to late 1970s in Boston. Using the results to infer information about other times and locations may or may not be appropriate. The study designers might hope that Boston is representative of much of the United States in terms of smoking among youth, but generalizing the results to other countries with different lifestyles or to the present day may be doubtful.

The youths in the FEV study are aged from 3 to 19. As no data exists outside this age range, no statistical model can be verified to apply outside this age range. In the same way, no statistical model applies for youth under 46 inches tall or over 74 inches tall. FEV cannot be expected to increase linearly for all ages and heights. □

1.13 Using R for Statistical Modelling

A computer is indispensable in any serious statistical work for performing the necessary computations (such as estimating the values of β_j), for producing graphics, and for evaluating the final model.

Although the theory and applications of GLMs discussed throughout this book apply generally, the implementation is possible in various statistical computer packages. This book discusses how to perform these analyses using R (all computations in this book are performed in R version 3.4.3). A short introduction to using R is given in Appendix A (p. 503).

This section summarizes and collates some of the relevant R commands introduced in this chapter. For more information on some command `foo`, type `?foo` at the R command prompt.

- `library()`: Loads extra R functionality that is contained in an R package. For example, use `library(GLMsData)` to make the data frames associated with this book available in R. See Appendix B (p. 525) for information about obtaining and installing this package.
- `data()`: Loads data frames.
- `names(x)`: Lists the names of the variables in the data frame `x`.
- `summary(object)`: Produces a summary of the variable `object`, or of the data frame `object`.
- `factor(x)`: Declares `x` as a factor. The first input is the variable to be declared as a factor. Two further inputs are optional. The second (optional) input `levels` is the list of the levels of the factor; by default the levels of the factor are sorted by numerical or alphabetical order. The third (optional) input `labels` gives the labels to assign to the levels of the factor in the order given by `levels` (or the order assumed by default).

- `relevel(x, ref)`: Changes the reference level for factor `x`. The first input is the factor, and the second input `ref` is the level of the factor to use as the reference level.
- `plot()`: Plots data. See Appendix A.3.10 (p. 516) for more information.
- `legend()`: Adds a legend to a plot.

1.14 Summary

Chapter 1 introduces the idea of a statistical model. In this context, y refers to the response variable, n to the number of observations, and $x_1, x_2, \ldots, x_p$ to the p explanatory variables. Quantitative explanatory variables are called covariates; qualitative explanatory variables are called factors (Sect. 1.2). Factors must be *coded* numerically for use in statistical models (Sect. 1.4) using dummy variables. Treatment codings are commonly used, and are used by default in R. $k-1$ dummy variables are required for a factor with k levels.

Plots are useful for an initial examination of data (Sect. 1.3), but statistical models are necessary for better understanding. Statistical models explain the two components of data: The *systematic component* models how the mean response changes as the explanatory variables change; the *random component* models the variation of the data about the mean (Sect. 1.5). In this way, statistical models represent both the systematic and random components of data (Sect. 1.8), and can be used for prediction, and for understanding relationships between variables (Sect. 1.9). Two criteria exist for an adequate model: simplicity and accuracy. The simplest model that accurately describes the systematic component and the randomness is preferred (Sect. 1.10).

Regression models 'linear in the parameters' have a systematic component of the form $\text{E}[y_i] = \mu_i = f(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})$ (Sect. 1.6). In these models, the number of regression parameters is denoted p'. If a constant term β_0 is in the systematic component, as is almost always the case, then $p' = p + 1$; otherwise $p' = p$ (Sect. 1.6).

Statistical models should be able to be sensibly interpreted (Sect. 1.7). However, fitted models should be interpreted and understood within the limitations of the data and of the model (Sect. 1.11). For example: in observational studies, data are simply observed, and no cause-and-effects conclusions can be drawn. In experimental studies, data are produced when the researcher has some control over the values of at least some of the explanatory variables to use; cause-and-effect conclusions may be drawn (Sect. 1.11). In general, conclusions from fitting and analysing a statistical model only apply to the population represented by the sample (Sect. 1.12).

Computers are invaluable in statistical modelling, especially for estimating parameters and graphing (Sect. 1.13).

Problems

Selected solutions begin on p. 529.

1.1. The plots in Fig. 1.7 (data set: `paper`) show the strength of Kraft paper [7, 8] for different percentages of hardwood concentrations. Which systematic component, if any, appears most suitable for modelling the data? Explain.

1.2. The plots in Fig. 1.8 (data set: `heatcap`) show the heat capacity of solid hydrogen bromide y measured as a function of temperature x [6, 16]. Which systematic component, if any, appears best for modelling the data? Explain.

1.3. Consider the data plotted in Fig. 1.9. In the panels, quadratic, cubic and quartic systematic components are shown with the data. Which systematic component appears best for modelling the data? Explain.

The data are actually randomly generated using the systematic component $\mu = 1 + 10\exp(-x/2)$ (with added randomness), which is not a polynomial at all. Explain what this demonstrates about fitting systematic components.

1.4. Consider the data plotted in Fig. 1.10 (data set: `toxo`). The data show the proportion of the population y testing positive to toxoplasmosis against the annual rainfall x for 34 cities in El Salvador [5]. Analysis suggests a cubic model fits the data reasonably well (though substantial variation still exists). What important features of the data are evident from the plot? Which of the plotted systematic components appears better? Explain.

1.5. For the following systematic components used in a regression model, determine if they are appropriate for regression models linear in the parameters, linear regression models, and/or generalized linear models. In all cases, β_j refers to model parameters, μ is the expected value of the response variable, while x, x_1 and x_2 refer to explanatory variables.

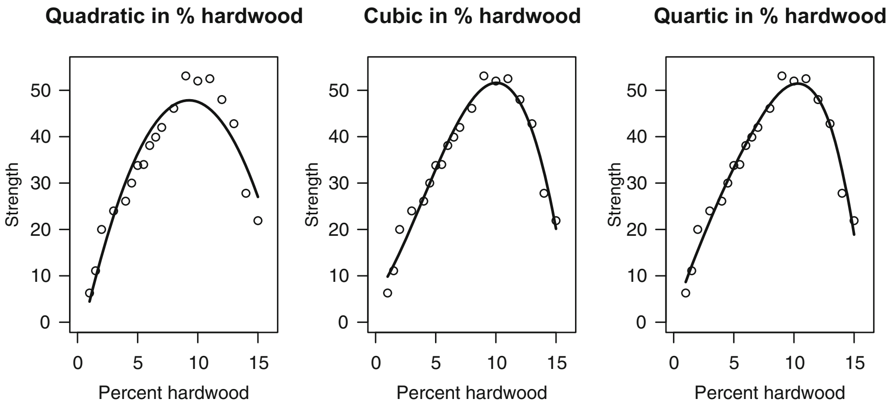

Fig. 1.7 Three different systematic components for the Kraft paper data set: fitted quadratic, cubic and quartic systematic components are shown (Problem 1.1)

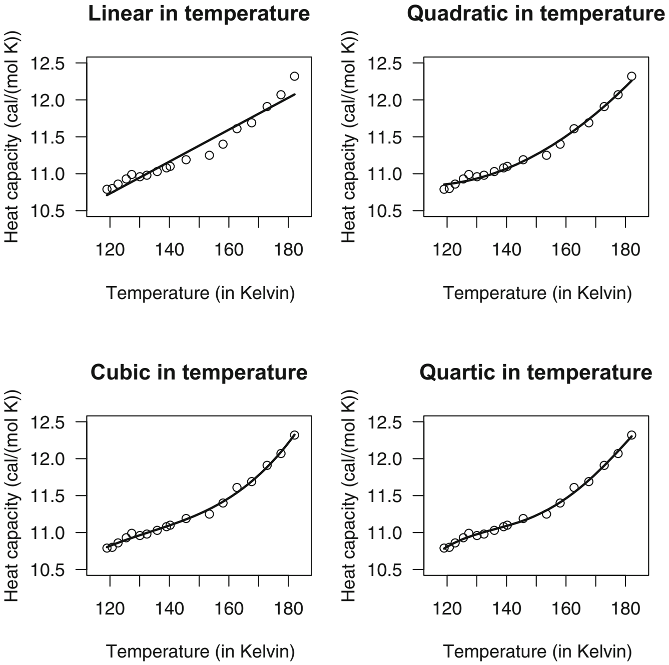

Fig. 1.8 Plots of the heat capacity data: fitted linear, quadratic, cubic and quartic systematic components are shown (Problem 1.2)

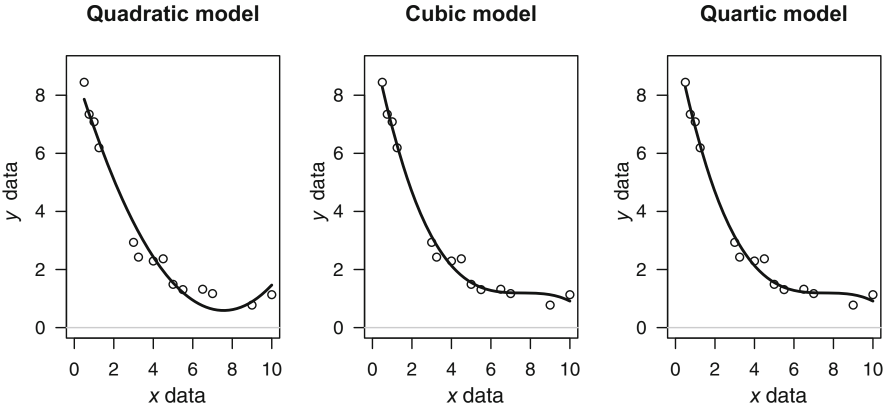

Fig. 1.9 Three different systematic components for a data set: fitted quadratic, cubic and quartic systematic components are shown (Problem 1.3)

1. $\mu = \beta_0 + \beta_1 x_1 + \beta_2 \log x_2$.
2. $\mu = \beta_0 + \exp(\beta_1 + \beta_2 x)$.
3. $\mu = \exp(\beta_0 + \beta_1 x)$ for $\mu > 0$.
4. $\mu = 1/(\beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2)$ for $\mu > 0$.

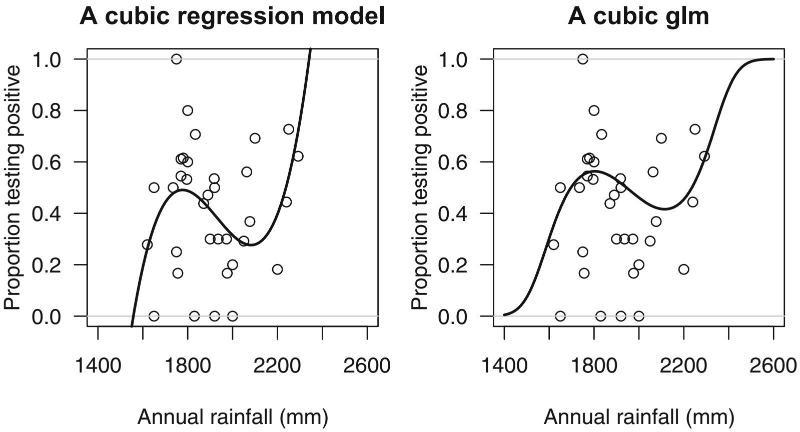

Fig. 1.10 The toxoplasmosis data, and two fitted cubic systematic components (Problem 1.4)

1.6. Load the data frame `turbines` from the package **GLMsData**. Briefly, the data give the proportion of turbines developing fissures after a given number of hours of run-time [13, 14].

1. Use `names()` to determine the names of the variables in the data frame.
2. Determine which variables are quantitative and which are qualitative.
3. For any qualitative variables, define appropriate dummy variables using treatment coding.
4. Use R to summarize each variable.
5. Use R to create a plot of the proportion of failures (turbines with fissures) against run-time.
6. Determine the important features of the data evident from the plot.
7. Would a linear regression model seem appropriate for modelling the data? Explain.
8. Read the help for the data frame (use `?turbines` after loading the **GLMsData** package in R), and determine whether the data come from an observational or experimental study, then discuss the implications.

1.7. Load the data frame `humanfat`. Briefly, the data record the percentage body fat y, age, gender and body mass index (BMI) of 18 adults [12]. The relationship between y and BMI is of primary interest.

1. Use `names()` to determine the names of the variables in the data.
2. Determine which variables are quantitative and which are qualitative. Identify which variables are extraneous variables.
3. For any qualitative variables, define appropriate dummy variables using treatment coding.
4. Use R to summarize each variable.

5. Plot the response against each explanatory variable, and discuss any important features of the data.
6. Would a linear regression model seem appropriate for modelling the data? Explain.
7. Read the help for the data frame (use `?humanfat` after loading the **GLMsData** package in R), and determine whether the data come from an experiment or observational study. Explain the implications.
8. After reading the help, determine the population to which the results can be expected to generalize.
9. Suppose a linear regression model was fitted to the data with systematic component $\mu = \beta_0 + \beta_1 x_1$, where x_1 is BMI. Interpret the systematic component of this model.
10. Suppose a generalized linear model was fitted to the data with systematic component $\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is BMI, and x_2 is 0 for females and 1 for males. Interpret the systematic component of this model.
11. For both models given above, determine the values of p and p'.

1.8. Load the data frame `hcrabs`. Briefly, the data give the number of male satellite crabs y attached to female horseshoe crabs of various weights (in g), widths (in cm), colours and spine conditions [1, 3].

1. Determine which variables are quantitative and which are qualitative.
2. For any qualitative variables, define appropriate dummy variables using treatment coding.
3. Use R to summarize each variable.
4. Produce appropriate plots to help understand the data.
5. Find the correlation between weight and width, and comment on the implications.
6. Read the help for the data frame (use `?hcrabs` after loading package **GLMsData** in R), and determine whether the data come from an experiment or observational study. Explain the implications.
7. After reading the help, determine the population to which the results can be expected to generalize.
8. Suppose a linear regression model was fitted to the data with systematic component $\mu = \beta_0 + \beta_1 x_1$, where x_1 is the weight of the crab. Interpret the systematic component of this model. Comment on the suitability of the model.
9. Suppose a generalized linear model was fitted to the data with systematic component $\log \mu = \beta_0 + \beta_1 x_1$, where x_1 is the weight of the crab. Interpret the systematic component of this model. Comment on the suitability of the model.
10. For the model given above, determine the values of p and p'.

1.9. Children were asked to build towers as high as they could out of cubical and cylindrical blocks [9, 17]. The number of blocks used and the time taken were recorded.

1. Load the data frame `blocks` from the package **GLMsData**, and produce a summary of the variables.
2. Produce plots to examine the relationship between the *time* taken to build towers, and the block type, trial number, and age.
3. In words, summarize the relationship between the four variables.
4. Produce plots to examine the relationship between the *number* of blocks used to build towers, and the block type, trial number, and age.
5. Summarize the relationship between the four variables in words.

1.10. In a study of foetal size [15], the mandible length (in mm) and gestational age for 167 foetuses were measured from the 15th week of gestation onwards. Load the data frame `mandible` from the package **GLMsData**, then use R to create a plot of the data.

1. Determine the important features of the data evident from the plot.
2. Is a linear relationship appropriate? Explain.
3. Is a model assuming constant variation appropriate? Explain.

References

[1] Agresti, A.: An Introduction to Categorical Data Analysis, second edn. Wiley-Interscience (2007)

[2] Box, G.E.P., Draper, N.R.: Empirical Model-Building and Response Surfaces. Wiley, New York (1987)

[3] Brockmann, H.J.: Satellite male groups in horseshoe crabs, *limulus polyphemus*. Ethology **102**, 1–21 (1996)

[4] Dunn, P.K., Smyth, G.K.: GLMsData: Generalized linear model data sets (2017). URL https://CRAN.R-project.org/package=GLMsData. R package version 1.0.0

[5] Efron, B.: Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association **81**(395), 709–721 (1986)

[6] Giauque, W.F., Wiebe, R.: The heat capacity of hydrogen bromide from 15°K. to its boiling point and its heat of vaporization. The entropy from spectroscopic data. Journal of the American Chemical Society **51**(5), 1441–1449 (1929)

[7] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.Y., Ostrowski, E.: A Handbook of Small Data Sets. Chapman and Hall, London (1996)

[8] Joglekar, G., Scheunemyer, J.H., LaRiccia, V.: Lack-of-fit testing when replicates are not available. The American Statistician **43**, 135–143 (1989)

[9] Johnson, B., Courtney, D.M.: Tower building. Child Development **2**(2), 161–162 (1931)

[10] Kahn, M.: An exhalent problem for teaching statistics. Journal of Statistical Education **13**(2) (2005)

[11] Maron, M.: Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation **136**, 100–107 (2007)

[12] Mazess, R.B., Peppler, W.W., Gibbons, M.: Total body composition by dualphoton (^{153}Gd) absorptiometry. American Journal of Clinical Nutrition **40**, 834–839 (1984)

[13] Myers, R.H., Montgomery, D.C., Vining, G.G.: Generalized Linear Models with Applications in Engineering and the Sciences. Wiley, Chichester (2002)

[14] Nelson, W.: Applied Life Data Analysis. Wiley Series in Probability and Statistics. John Wiley Sons, New York (1982)

[15] Royston, P., Altman, D.G.: Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Journal of the Royal Statistical Society, Series C **43**(3), 429–467 (1994)

[16] Shacham, M., Brauner, N.: Minimizing the effects of collinearity in polynomial regression. Industrial and Engineering Chemical Research **36**, 4405–4412 (1997)

[17] Singer, J.D., Willett, J.B.: Improving the teaching of applied statistics: Putting the data back into data analysis. The American Statistician **44**(3), 223–230 (1990)

[18] Smyth, G.K.: Australasian data and story library (OzDASL) (2011). URL http://www.statsci.org/data

[19] Tager, I.B., Weiss, S.T., Muñoz, A., Rosner, B., Speizer, F.E.: Longitudinal study of the effects of maternal smoking on pulmonary function in children. New England Journal of Medicine **309**(12), 699–703 (1983)

[20] Tager, I.B., Weiss, S.T., Rosner, B., Speizer, F.E.: Effect of parental cigarette smoking on the pulmonary function of children. American Journal of Epidemiology **110**(1), 15–26 (1979)